The Evolution of Data Workforce Requirements in Science and Engineering Libraries

Jeonghyun Kim[1], Sujira Ammarukleart[1]
[1]University of North Texas

Abstract
As data-intensive science has emerged, researchers are expected to discover, collect, process, analyze, archive, and share data in their everyday lives and the need for developing proficiency with research data has been recognized. It has been asserted that information professionals in the fields of science and engineering who have supported researchers through a variety of activities are well positioned to assist and meet such needs. This study applied a text mining approach to investigate changes in the requirements for information professionals in science and engineering, with a particular focus on duties and responsibilities to support research data stewardship. Position advertisements posted in the Association of College & Research Libraries' Science and Technology Section Discussion List from 2010 to 2014 were collected and analyzed.

Keywords: Job advertisements; job analysis; text mining; research data stewardship
doi: 10.9776/16511

Contact: Jeonghyun.Kim@unt.edu

1 Introduction

The exponential growth in the amount of data has become a big topic across nearly every area of information technology and has led to the development of a new competitive arena as revolutionary measures are needed for the management, analysis, and accessibility of big data. Such data growth has fundamentally changed the landscape of scientific research as well. We are now in the "fourth paradigm" (Gray, 2009) of data-intensive science, where "all of the science literature is online, all of the science data is online, and they interoperate with each other" (Howe et al., 2008, p. 47). This new paradigm has gained significance in cutting-edge scientific research. In light of this movement, the importance of curating research data, which refers to the management of data throughout its research lifecycle, has been highlighted. The sustainability of the long-term access and preservation of research data is a well-known challenge and requires an institutional level of support. To this end, academic libraries have found themselves embracing new roles to provide institutional responses for these challenges. This is evidenced by the Association of the College & Research Libraries (ACRL), which identified library involvement in research data services as one of 2015's top trends in academic libraries (ACRL Research Planning and Review Committee, 2015).

Along with discussion on re-envisioning research and academic libraries' role, increased interest has been sparked in broadening the role of information professionals. In many disciplines, such as science, business, and health, information professionals are paying attention to new professional demands and publishing studies on the relevance and implications of data management in their specific fields. For instance, a report, *Transforming Liaison Roles in Research Libraries*, released by the Association of Research Libraries (ARL), claimed that subject specialists can make a significant contribution to that area as they bring their understanding of the research methods in their assigned disciplines, considering the size and formats of the data produced, and the availability of disciplinary repositories (Jaguszewski & Williams, 2013). As such, there has been a strong call for information professionals to play a role in the broad area of research data. Among various disciplines, the fields of science and engineering are experiencing an unprecedented data avalanche due to the fast advance and evolution of information technology that enables capture, analysis, and storage of huge quantities of data.

The purpose of this study is to address the following question: Are the duties and responsibilities of information professionals in the fields of science and engineering changing and expanding to support research data stewardship?

2 Methodology

To educate the future professional workforce and meet the demand for new professionals in the field, much attention needs to be paid to the ever-changing skills and knowledge expectations in the labor market. Job advertisements, which include rich information about the requirements, knowledge, skills, abilities, education, and experience considered essential or desirable in the individual who will fill the position, have been studied in library and information science. To analyze such job advertisements, content analysis has been regarded as an established methodology in the discipline. However, the amount of human effort required to collect, transcribe, and code textual data made content analysis time-consuming and labor-intensive.

In recent years, text mining has been adopted to discover trends in textual data as it has been regarded suitable for drawing valuable information from large volumes of unstructured text. *N*-gram analysis is one of the leading text-mining techniques commonly used in the field of computational linguistics to analyze a large corpus of documents with the purpose of identifying combinations of keywords that frequently appear together (Manning & Schtze, 1999). This approach does more than a simple frequency analysis because it accounts for how words are combined. Moreover, if the documents are ordered in time, it is feasible to do a trend analysis on the popularity of certain keywords. This study applied an *n*-gram analysis approach to investigate changes in the duties and responsibilities for the position in the science and engineering libraries.

2.1 Data Collection

Job advertisements were harvested from the Association of College & Research Libraries' Science & Technology Section Discussion List (http://lists.ala.org/wws/info/sts-l), which has served as a forum for information professionals in scientific and technical subject fields to maintain awareness of the impact and range of information with which they work. Position announcements posted from 2010 to 2014 were collected to investigate changes in the science and technology information professional workforce. The year 2010 was chosen as a starting point as the discussion on the role of information professionals in eScience was initiated in the late 2000s, and many major funding agencies, such as National Science Foundation, mandated data sharing and management plans in 2010.

A total of 321 position announcements were collected. However, duplicate announcements and announcements that did not include job qualification sections were excluded. As such, a total of 294 announcements were included in this study for analysis.

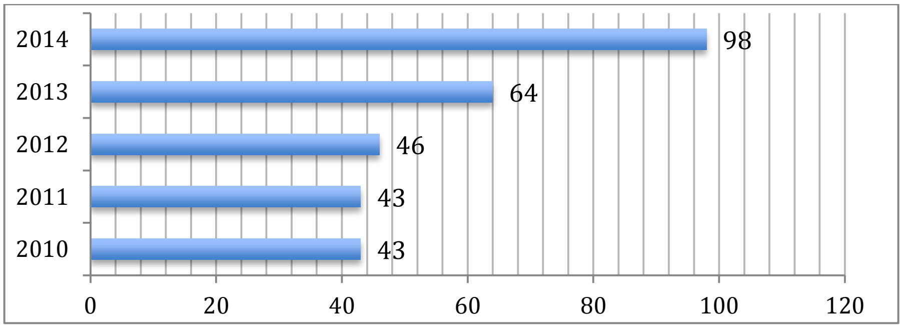

Figure 1. Job announcements from 2010 to 2014

2.2 Data Analysis

Each job announcement was examined for position title, duties, and responsibilities. Titles were analyzed because job titles have a significant role in reflecting the changes in responsibilities. To analyze duties and responsibilities statements from each job advertisement, the sentences under the "Duties," "Responsibilities," "Essential functions," and/or "Key accountabilities," were first extracted. Additionally, the following sentences were taken from the description: "The successful candidate will…" "The position is responsible for…" and/or "Responsibilities (or duties) include…."

Rapidminer 5.3 was used to generate *n*-grams. For this study, we limited our analysis to include *n*-grams with a maximum length of n=3; as a result, a number of unigrams, bi-grams, and tri-grams were generated. The researchers compared the frequency with which every *n*-gram in the corpus appeared in job announcements posted during each year of analysis. However, direct frequency comparisons can be deceiving, since they do not account for the potential growth or decline in the number of words used in job titles over time. Therefore, it is necessary to calculate relative frequency for each *n*-gram. We divided *n*-gram frequencies (*n*-gram frequency was derived from job announcement frequency, which refers to the number of job announcements in the corpus that contain a particular term) for each year by the total number of jobs posted during that year in order to produce the standardized measure of frequency that would allow valid comparison between *n*-grams across years (i.e., Relative frequency = *n*-gram occurrences/total number of job announcements).

3 Results

3.1 Position Titles

The most frequently occurring unigram used in the position titles, “librarian,” was found in 81 percent of announcements. Figure 2 presents the trends of the top 10 unigrams that experienced the most growth or decline in popularity in the position titles from 2010 to 2014.

The terms “research” (+19.84%), “services” (+10.09%), and “data” (+7.59%) show the most growth in the position titles; these position titles include bi-grams such as “research data,” “data curation,” “data librarian,” and “research services.” This implies that the increasing positions advertised reflect the role in supporting research data stewardship. However, the terms “science” (-16.59%) and “reference” (-4.49%) declined steadily from 2010 to 2014. It should be noted that the term “science” was often used to denote a discipline-specific description, such as life science, health science, and physical science, but it became less common. It is also worthwhile to note that the term “eScience” had grown until 2013, but there were no positions including the term “eScience” in 2014.

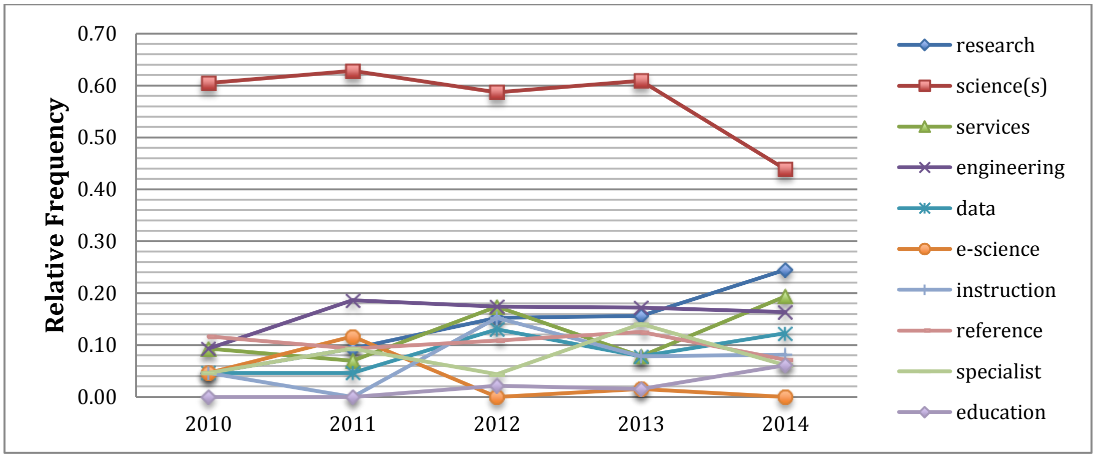

Figure 2. Trends in unigram used in the position titles

3.2 Duties and Responsibilities

Figure 3 presents the trends of top 10 bi-grams that experienced the most growth or decline in popularity in the job duties and responsibilities statements from 2010 to 2014.

Among the top five bi-grams with highest popularity growth were “data management” (+20.57%), “research data” (+17.80%), “data service(s)” (+12.70%), and “research service(s)” (+7.76%); this indicates that services for scientific research and research data as a core duty are increasing. More positions are expected to serve a key role in the expansion of a library’s research data management practices and services. This is also confirmed by lists of bi-grams and tri-grams for research and research data-related duties appearing from the job announcements, as presented in Table 1. An example of a statement that appeared in these job ads is a scientific data curator who will “facilitate the collection, preservation, and access to scientific data, and will act as a resource for students and faculty grappling with issues of data curation, digital methods for scientific research, and emerging digital resources.” Another example includes an engineering and informatics librarian who should “develop and support services for documenting and distributing research data and develops and maintains expertise in data issues for libraries.”

Libraries are also asking information professionals to provide instructional sessions and consultations to meet users’ research needs. This is evidenced by the growth of bi-grams such as “research consultation” (+10.65%) and “instruction research” (+8.61%). The decrease of the bi-grams “library instruction” (-6.36%) and “classroom instruction” (-16.28%) also implies that more instruction and consulting are geared toward research for faculty, graduate students, and/or research teams rather than curriculum support or general database instruction. Some emerging instruction and consultation duties include providing training for users in data management best practices and standards, offering instruction and consultation to faculty and graduate students in response to information needs related to grant preparation, data management, data sharing, and publication, and developing/delivering training and instructional materials on data curation. Additionally, both in-person and virtual reference assistance is part of this continuing commitment.

Collection development and management for assigned subjects were still mentioned in the announcements without any noticeable growth or decline. However, it is apparent that more positions are expected to actively participate in users' research activities to meet their research needs through such traditional works. This is evidenced by the following statements: "Building collections in all formats that support research and instructional needs," and "Collaborating with Collection and System Management to insure data licensing and procurement needs of the campus community are met."

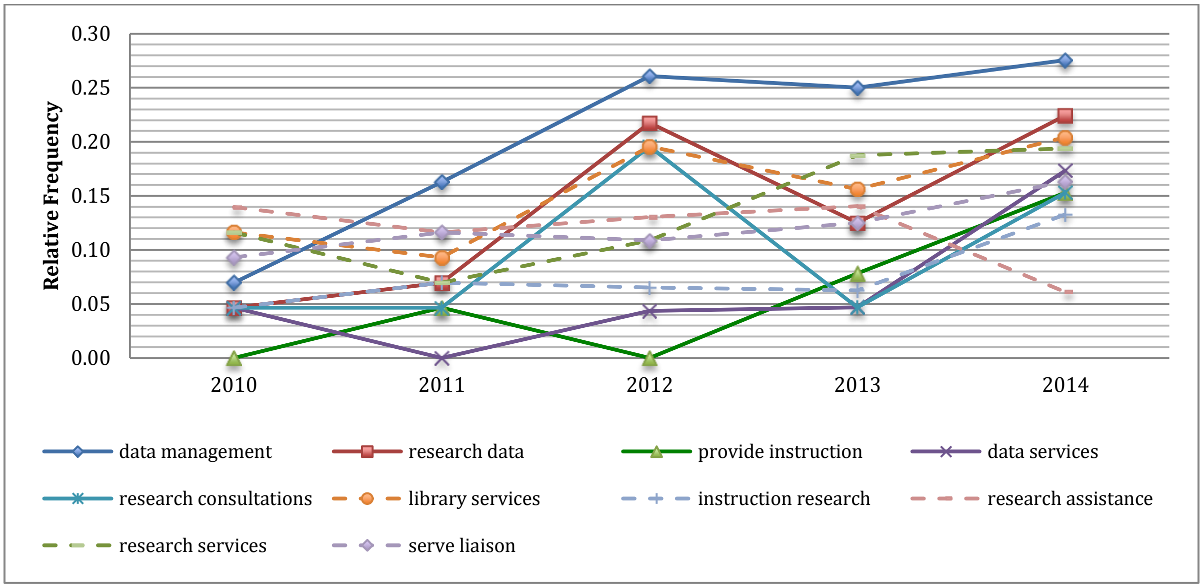

Figure 3. Trends in bi-grams used in the duties/responsibilities statements

Year	Bi-grams and tri-grams appearing	Number of Jobs	Percentage
2010	data curation, data management, data preservation, data services, reference research, research assistance, research support, research teaching, research services, research data	22	51.16%
2011	data management, data management plans, data management planning, data curation reference research, research assistance, research support, research teaching, assigned lab research, lab research discovery, provide research support, research consultation	21	48.84%
2012	data management, data curation, data management plans, issues data curation, data librarian, reference research, research data, research consultation, research support, support research, research services, research assistance depth reference research	29	63.04%
2013	data management, data curation, data management plans, data management curation, management curation, data curation data, curation data, curation visualization, data mining, data services, research services, research assistance, reference research, research data, research instruction, teaching research, research support, research guides, subject specific research, research consultations	43	67.19%
2014	data management, data services, research data management, data curation, data management plans, curation preservation, data issues, management curation, data management curation, curation research, curation specialist, data curation specialist, data sharing, research data, research service, research support, research consultation, reference research, research support services, research assistance, research guides	66	67.35%

Table 1. Bi-grams and tri-grams for research-related duties

4 Conclusions

This paper reports the preliminary findings of an analysis of job announcements to explore how information professionals' roles and responsibilities in science and engineering have evolved to meet data challenges. The reported results imply that there has been a call for developing services across the research lifecycle model from planning a research proposal to publishing findings and sharing data.

We are currently carrying out further analyses; the results of such analyses will be presented at the conference.

5 References

ACRL Research Planning and Review Committee. (2015). *Environmental scan 2015*. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/EnvironmentalScan15.pdf

Gray, J. (2009). Jim Gray on eScience: A transformed scientific method. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The fourth paradigm: Data-intensive scientific discovery*, (pp. xvii-xxxi). Redmond, WA: Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf

Jaguszewski, J. M., & Williams, K. (2013). *New roles for new times: Transforming liaison roles in research libraries*. Retrieved from http://www.arl.org/storage/documents/publications/NRNT-Liaison-Roles-final.pdf

Howe, D., et al. (2008). Big data: The future of biocuration. *Nature, 455*(7209), 47-50.

Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.